DOCUMENT RESUME

ED 194 612 TM 800 743

AUTHOR	Brossell, Gordon; Hoetker, James
TITLE	Examination Handbook for the Writing Subtest of the Florida Teacher Competency Examination. Volume One of Five.
INSTITUTION	Florida State Univ., Tallahassee. Coll. of Education.
SPONS AGENCY	Florida State Dept. of Education, Tallahassee.
PUB DATE	Jun 79
NOTE	48p.: For related documents, see TM 800 744-47.
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Competency Based Teacher Education; Cutting Scores; Elementary Secondary Education; *Essay Tests; Evaluation Criteria; Evaluation Methods; Evaluators; Higher Education; *Minimum Competency Testing; Student Evaluation; Teacher Certification; Test Construction; Test Format; Writing (Composition); *Writing Skills
IDENTIFIERS	Florida; *Florida Teacher Competency Examination; *Holistic Evaluation; Interrater Reliability; Writing Evaluation

ABSTRACT

The writing examination package is divided into three volumes: Examination Handbook (present volume), Training Manual, and a Ratings Manual. The volumes are interdependent, and all three must be understood before using the Training Manual to train raters. The Examination Handbook provides a context for the other two volumes. It gives a general description and background of the holistic approach to evaluating examination essays, and provides discussion of matters such as test preparation, rater selection, reliability computations, developmental uses of data from the first administration of the examination, and possible research uses of the examination essays and ratings. Major recommendations include that: (1) the form of holistic evaluation to be used is "general impression marking;" (2) the assignment format consist of at least six optional topics; (3) the examinees be given a minimum of 45 minutes to write the essay; (4) teams of 3 raters, all having backgrounds as high school writing teachers, be used to score essays; and (5) cutoff score be "5" on a scale running from "3" to "12." (Author/RL)

EXAMINATION HANDBOOK FOR THE WRITING SUBTEST OF THE FLORIDA TEACHER COMPETENCY EXAMINATION

Gordon Brossell James Hoetker

College of Education
Florida State University

VOLUME ONE OF FIVE

Under Contract to the Department of Teacher Education
Florida State Department of Education

#790116

Department of Education

June, 1979

ACKNOWLEDGMENTS

We wish to thank the following people who assisted us in the preparation of these documents: Mrs. Lucy Harward, who conducted the field trials of the materials and served as our general editor and severest critic; Professors Nancy McGee (University of Central Florida) and Dan Kelly (University of Florida), who graciously reviewed the entire package, making innumerable valuable suggestions; and, finally, Miss Theresa Pistana, who patiently and good-naturedly typed the several versions of this manuscript.

WRITING EXAMINATION HANDBOOK TABLE OF CONTENTS

1. DESCRIPTION OF THE WRITING EXAMINATION PACKAGE.

The writing examination package is divided into three volumes, an EXAMINATION HANDBOOK (the present volume), a TRAINING MANUAL, and a RATINGS MANUAL. The volumes are interdependent and it is important that one be familiar with all three volumes before, for instance, attempting to use the Training Manual to train raters. The contents of the volumes are as follows.

A. The EXAMINATION HANDBOOK provides a context for the other two volumes. It gives a general description and background of the holistic approach to evaluating examination essays, and provides discussion of matters such as test preparation, rater selection, reliability computations, and so on.

B. The TRAINING MANUAL is addressed to the individuals in charge of training the raters and referees who will evaluate the essays written by the examinees. It consists of two parts:

   a. A trainer's guide which lays out, in sequential "lesson plan" form, the steps in the training process; and

   b. A packet of materials to be duplicated for the raters, consisting of instructions, criteria for rating, sample student essays, and so on.

C. The RATINGS MANUAL is addressed to the administrator and the clerk(s) who will be responsible for the smooth conduct of the process of rating the student essays. It consists of three parts:

a. A detailed chronological description of the steps in the ratings process, into which are inserted discussions of various matters directly pertinent to the process--estimating manpower needs, assigning essays to rater teams, and so on.

b. An assemblage of copies of the various forms that will be used in the ratings process, the forms being completed step by step to illustrate the processes discussed in the chronological description.

c. An assemblage of blank copies of the forms to be used in the ratings process, to be duplicated for use as prescribed.

2. METHODS OF EVALUATING WRITING ABILITY

The Council on Teacher Education recommendations for the writing subtest of the Teacher Competency Examination specify that it be "a writing production test that will be rated holistically by selected evaluation experts."

A writing production test is one of two basic methods of obtaining a measure of someone's writing ability. It might be called the "direct" method, in that it involves rating directly a sample of writing. The other--"indirect"--method is to administer an objective test of some trait that is ostensibly related to writing ability. Examiners using the indirect approach have sampled such things as knowledge of grammar and usage rules, ability to recognize errors and edit a flawed passage, range of vocabulary, and verbal reasoning ability. Testmakers have presented evidence that a carefully constructed objective test can be a highly valid predictor of writing ability. The conviction still persists, however, especially among teachers of writing, that no test that does not involve the production of writing can really be called a test of writing ability.

Two methodologies for directly evaluating the quality of essays have been developed--the analytical and the holistic. In the analytical approach, the rater, guided by some sort of essay scale or checklist of essay characteristics, reads an essay as many times as necessary for him to make a judgment of the quality of the essay in regard to each of the characteristics identified on the checklist (e.g., organization, style, vocabulary, mechanics, syntax, spelling, etc). The rater will commonly award a number score on each

characteristic, with the total of those scores being the grade for the essay. This sort of approach is time-consuming and therefore expensive, and is more appropriate for research and diagnostic purposes than for a simple assessment of quality. In the holistic approach, several readers read an essay only once to form a general impression of its quality, or for some more specific purpose. Each awards the essay a rating indicative of his or her judgment of it; and the sum or average of their ratings is the score for the essay.

As the problem with an objective test of writing ability is its validity, so the problem with a writing production test is its reliability or consistency. (Reliability may be defined roughly as the probability that an essay will be awarded the same grade again if the evaluation procedure is repeated.) Although analytical and holistic ratings of essays are subjective, many years of work grading essay examinations have demonstrated that if there are multiple readers, and if the readers are carefully trained, very high inter-rater agreement can be obtained.

Perhaps the best non-technical discussion of the whole matter of evaluating production tests of writing ability is Measuring Growth in English (Urbana, IL: NCTE, 1974) by Paul Diederich, who has pioneered in the development of both analytical and holistic methods of evaluation at the Educational Testing Service. Cooper and Odell's Evaluating Writing (Urbana, IL: NCTE, 1977) provides thorough discussions of the state-of-the-art in a variety of direct approaches to assessing writing skills. Foley's review of the literature in "Evaluation of Learning in Writing," in Bloom, et al.,

Formative and Summative Evaluation of Student Learning (New York: McGraw-Hill, 1971) contains references to the major research studies. A College Entrance Examination Board pamphlet, "Guide to Examinations in English" (Princton, NJ, 1974), succinctly describes the holistic approach to evaluation and reports on reliabilities obtained by CEEB. Roberts and Rentz have edited a collection of papers, "Research Related to the Reliability and Validity of the Language Skills Examination of the Regents' Testing Program" (Atlanta, GA: Georgia State University, mimeographed, 1978), which are especially pertinent to the Teacher Competency Examination. An interesting brief history of the College Board's attempt to measure writing ability--starting in 1901 with 2-1/2 and 3 hour essay examinations--may be found in Harris' "The Testing of Student Writing Ability," in Tate, ed., Reflections on High School English (Tulsa, OK: University of Tulsa Press, 1966).

3. THE GENERAL IMPRESSION METHOD OF HOLISTIC EVALUATION AND COTE'S ESSENTIAL WRITING COMPETENCIES

A. The "General Impression" Approach to Evaluation

In his essay on "Holistic Evaluation of Writing" (in Cooper and Odell, Evaluating Writing, pp. 3-31), Charles Cooper gives this general definition of the procedure:

> Holistic evaluation of writing is a guided procedure for sorting or ranking written pieces. The rater takes a piece of writing and either (1) matches it with another

> piece in a graded series...or (2) scores it for the prominence of certain features...or (3) assigns it a letter or number grade. The placing, scoring, or grading occurs quickly, impressionistically, after the rater has practiced the procedure with other raters. The rater does not make corrections or revisions in the paper. Holistic evaluation is usually guided by a holistic scoring guide which describes each feature and identifies high, middle, and low quality levels for each feature.... Holistic evaluation remains the most valid and direct means of rank-ordering students by writing ability. Spending no more than two minutes on each paper, raters... can achieve a scoring reliability as high as .90 for individual writers. (p. 3)

The particular type of holistic evaluation employed for this examination is called "general impression marking," and it assigns number grades (or--the term employed in this document--"ratings") to the examinees' essays.

In this approach, again according to Cooper, "The rater simply scores the paper by deciding where the paper fits within a range of papers produced for that assignment." (pp. 11-12)

> As this procedure has been developed by Education Testing Service and the College Entrance Examination Board..., raters must train themselves carefully--become "calibrated" to reach consensus--by reading and discussing large numbers of papers like those they will be scoring. (p. 12)

Often, discussions of essays are guided by lists of criteria of quality. But even when no list of criteria is used, if raters are given the opportunity to discuss many papers, high inter-rater agreement has commonly been achieved, so that it may be assumed that the raters have developed an "implicit list of features or qualities to guide their judgment." (p.12)

The training procedures for the raters of the Teacher Competency Examination--as detailed in Volume Two of these materials--make use of both a detailed set of criteria and an extended period of guided discussion in order to assist the raters in internalizing a common set of "features or qualities to guide their judgment."

B. Description of Rating Procedures Used With This Examination

After the training session, the raters will be divided into teams of three. Each member of a team will read all the essays assigned to that team. A rater will read each essay quickly and only once and assign it a rating signifying his or her judgment of its quality. The ratings will range from "1" for "unsatisfactory" (or non-mastery) up to "4" for "outstanding"--so that ratings of "2," "3," and "4" all will signify mastery of writing skills at an acceptable level.

If the three raters do not agree with one another to the extent that their ratings are not confined to adjacent scores--that is, if any one of the ratings differs from another by two or more--then the essay will be forwarded to a referee or master rater for another reading. The referee's rating will replace the most discrepant of the original ratings.

The score awarded to an essay will be the sum of the ratings of three raters and may, therefore, range from "3" up to "12." A score of "5"--two raters awarding a passing grade, one a failing grade--will be the minimal passing score (see the discussion of the cut-off score below in Section 5).

C. Criteria for the Evaluation of Essays

The criteria according to which the raters of the Florida Teacher Competency Examination Subtest in writing will be trained must have two characteristics.

1. They must include those characteristics widely accepted as indicative of good writing; and

2. They must include those characteristics prescribed in COTE's listing of Essential Skills Competencies in Writing--that is, they must describe features of good writing that can reasonably be expected to be employed by college graduates seeking teacher certification in Florida.

For these purposes, the following criteria are submitted:

1. Rhetorical Quality

   1.1 Unity: An ordering and interdependence of parts producing a single effect; completeness.

   1.2 Focus: Concentration on the chosen topic.

   1.3 Clarity: Lucidity of expression; lack of ambiguity and distortion.

1.4 Sufficiency: Appropriate depth and breadth of expression to meet the writer's purposes and the demands of the particular topic.

2. Structural and Mechanical Quality

2.1 Organization: Consistent and coherent integration and connection of parts.

2.2 Development: Appropriate and sufficient exposition of ideas; use of detail, examples, illustrations, comparisons, etc.

2.3 Paragraph and Sentence Structure: Appropriate form, variety, logic, relatedness of and among structural units.

2.4 Syntax: Appropriate ordering of words to convey intended meaning.

3. Observance of Conventions in Writing

3.1 Usage: Appropriate use of language features: inflections, tense, agreement, pronouns, modifiers, vocabulary, level of discourse, etc.

3.2 Spelling, Capitalization, Punctuation: Consistent practice of accepted forms.

D. Operational Definitions of Levels of Quality

For purposes of rating, these criteria will be more useful to the raters if they are translated into four operational definitions corresponding to the four levels of writing competence.

This translation may be made as in the set of definitions below.

4. The essay is unified, sharply focussed, and distinctively effective. It treats the topic clearly, completely, and in suitable depth and breadth. It is clearly and fully organized, and it develops ideas with consistent appropriateness and thoroughness. The essay reveals an unquestionably firm command of paragraph and sentence structure. Syntactically, it is smooth and often elegant. Usage is uniformly sensible, accurate, and sure. There are very few, if any, errors in spelling, capitalization, and punctuation.

3. The essay is focussed and unified, and it is clearly if not distinctively written. It gives the topic an adequate though not always thorough treatment. The essay is well organized, and much of the time it develops ideas appropriately and sufficiently. It shows a good grasp of paragraph and sentence structure, and its usage is generally accurate and sensible. Syntactically, it is clear and reliable. There may be a few errors in spelling capitalization, and punctuation, but they are not serious.

2. The essay has some degree of unity and focus, but each could be improved. It is reasonably clear, though not invariably so, and it treats the topic with a marginal degree of sufficiency. The essay reflects some concern for organization and for some development of ideas, but neither is necessarily consistent nor fully realized. The essay reveals some sense, if not full command, or paragraph and sentence structure. It is syntactically bland and, at times, awkward. Usage is generally accurate, if not consistently so. There are some errors in spelling, capitalization, and punctuation

that detract from the essay's effect if not from its sense.

1. The essay lacks unity and focus. It is distorted and/or ambiguous, and it fails to treat the topic in sufficient depth and breadth. There is little or no discernible organization and only sporadically a sense of paragraph and sentence structure, and it is syntactically slipshod. Usage is irregular and often questionable or wrong. There are serious errors in spelling, capitalization, and punctuation.

E. How the Criteria Correspond to the Essential Competencies

The COTE phrasing of most of the subskill specifications allows a candidate for certification to demonstrate mastery of a subskill either indirectly--by answering a question requiring knowledge of the subskill--or directly by application of that subskill. As we have explained above, the holistic approach to evaluating a writing production test directly measures the candidate's ability to apply the essential writing competencies.

Figure 1 below shows graphically how the criteria that will be used to train the raters of the essays correspond to the list of essential competency subskills. Each of the subskills, it will be seen, is addressed in several of the criteria.

INSERT FIGURE 1 HERE

FIGURE 1. How the Essential Competency Subskills in Writing are Evaluated by a Criterion Guided Holistic Rating Procedure

ESSENTIAL COMPETENCIES: Demonstrate the ability to write in a logical, easily understood style with appropriate grammar and sentence structure.

A. Differentiate between formal and informal written English.

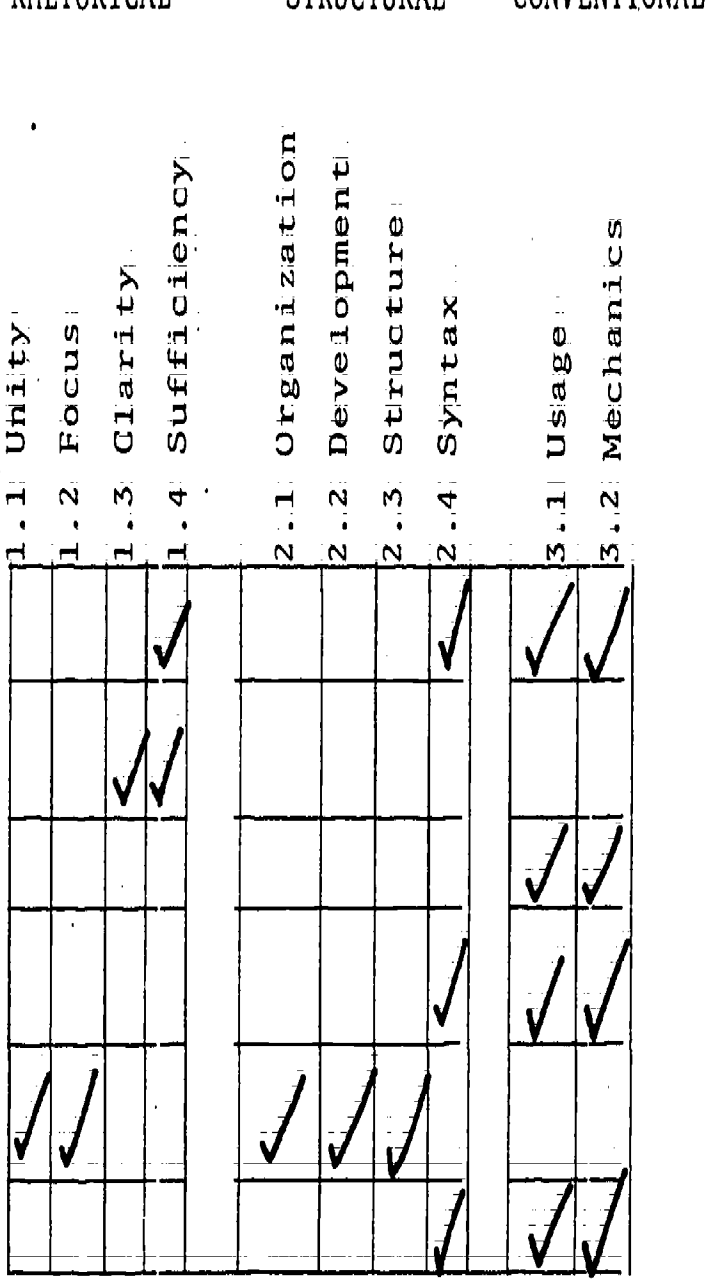
B. Use language appropriate to the topic and reader.

C. Apply basic mechanics of writing.

D. Apply appropriate sentence structure.

E. Apply basic techniques for organization.

F. Apply standard English usage.



4. WRITING THE EXAMINATION INSTRUCTIONS AND ASSIGNMENTS

A. Instructions to Examinees for Writing the Essay

The instructions should economically inform the examinee what he or she is expected to do and give him or her some information about how the essay will be evaluated. They should not--as instructions for such examinations sometimes seem to do--try to give the examinee a compressed course in how to write an essay; that is patronizing, intimidating, and wastes valuable time. The tone of the instructions should be friendly and supportive. Research recently reported by Michael Clark of the University of Michigan (at the Ottawa Conference on Learning to Write, May, 1979) demonstrates that the extraverbal features of writing test instructions may be as important as their content. Instructions, for instance, that are curt, peremptory, or harsh in tone may produce anxiety which interferes with the examinee's ability to concentrate and willingness to perform.

We suggest the following instructions as adequate, helpful, and non-threatening.

INSTRUCTIONS. This portion of the examination gives you a chance to show how well you can write. The question below asks you to compose an essay setting forth your personal opinions or beliefs on some important issue. You should assume you are addressing your essay to an audience of educated adults. Your purpose will be to convey your position as clearly as possible to your readers.

There are, of course, no "right answers" on this examination. Your essay will be read by at least three readers and judged on its quality as a prose composition. So use your time well--plan before you begin to write, then read your essay carefully after you have finished and make any necessary corrections and revisions. The evaluation of your essay will in no way depend on whether your readers happen to agree with your opinions. (But any reader will naturally appreciate legible handwriting.)

Relax, take a deep breath, and do the best you can.

B. Composing Assignments For the Writing Examination

There is almost no good research on the relationship between the type of assignment set on an examination of this sort and the quality of essays produced by the examinees. The recommendations made below, therefore, are made on the basis of logic and experience, and may be considered as testable hypotheses about the sort of stimuli that will produce the best writing of which an examinee is capable.

The purpose of an examination of writing skills is not to determine how much an examinee knows about some particular subject, but rather to determine how well he can express himself about (1) some subject with which he is already familiar, or (2) some proposition which calls for analysis and the application of principles rather than information. The good assignment, then, is one that identifies a topic or topics with which all of the examinees can reasonably be expected to be conversant or

able to handle without preparation.

Some authorities have at times claimed that the best topics for a writing examination are dull and trivial ones--topics such as, "How to tie a shoelace" or "How to drive a stick shift automobile." The reasoning is that the evaluation of essays written on such topics will be "pure" and unconfounded by rater reactions to the examinee's opinions or beliefs. We reject this position completely, on the grounds that (1) an examinee can do his best writing on a topic with which he feels some personal involvement and about which he has some genuine motivation to communicate, and (2) that rater boredom with one such dull paper after another would be a much greater threat to reliability than rater distraction by extreme opinions. The good assignment, then, should deal with an issue of some importance within the experience of the examinees.

The good assignment should also, obviously, be clearly and unambiguously phrased; should unequivocally inform the examinee just what sort of written product he is expected to turn out; and should specify a topic or topics of a "size" that can be dealt with in the allotted time.

Probably the commonest form taken by examinations of writing ability is that of a list of from six to ten optional topics from among which the examinee is to choose. Here for example is an assignment used in some research at Florida State University.

Read the topics below and choose one on which to write an essay.

1. Which person in public life do you most admire and why?
2. Explain what values you feel schools should impart to students.
3. In what ways does television affect you?
4. Explain why you favor or oppose the women's liberation movement.
5. What are the essential characteristics of a good teacher?
6. Should sex education be taught in American public schools or not?
7. Do viable alternatives to marriage exist in our society? Discuss.
8. Does your public image differ from your private self?

This form of assignment is time-honored, and there is little evidence that, if the topics are clearly stated, it is inferior to any other. There are any number of easily available sources of topics from which items may be borrowed or adapted. The National Council of Teachers of English, for example, publishes Grace E. Wilson's Composition Situations, in which hundreds of topics are organized in 45 categories, and distributes a leaflet descriptively entitled, A Thousand Topics for Composition.

One problem with this form of assignment, which CEEB has noted and which we have noticed in our own work, is that it produces essays in a wide variety of rhetorical modes--autobiographical reminiscences, arguments, editorials, meditations, lay sermons, and whatever. The raters in our work have reported they had problems adjusting to this melange of modes and found themselves applying

different standards to different kinds of discourse--to the detriment of inter-rater agreement.

Another approach to the writing examination involves devising a single broad assignment--sometimes an elaborately structured one--that is set for everyone to answer. The problem here is to find a topic that can be fair to all the examinees in a large and diverse population.[1]

The great advantage of the single set assignment is that it elicits rhetorically homogeneous responses--at least to the extent that the examinees answer the question that has been set. This is, we believe, of great advantage to the raters and should produce improved inter-rater reliability.

We are suggesting that the form of assignment used for this writing examination be one that combines the advantage of the topic-list--a variety of options--with the advantage of the single set topic--rhetorical homogeneity. Specifically, we suggest that the question take the form of a single set of directions associated with a set of optional topics cast in the same form. We suggest further, as already implied above, that the rhetorical mode prescribed by the assignment be that of the examinee expressing his own subjective opinions in his own voice. This seems to us likeliest to reduce anxiety and encourage fluency and freedom of expression.

---

[1] A third possible type of assignment would be an "open" one: e.g., "Choose an important educational issue and write an essay explaining your position on it." This approach has the fatal weakness that, once the word of the open format got out, examinees could "rehearse" their essays before taking the examination.

The assignment should be so structured that--in the context of the instructions presented in the preceding section which specify audience and purpose--it produces a writing situation that is clear, unambiguous, and (probably) familiar.

C. Sample Assignments and Topics

Here are three possible forms such an assignment might take. The first--which we would personally prefer--uses controversially-worded statements involving some issue of justice or equity as stimuli. The second uses direct questions as stimuli and allows for the presentation of topics that cannot be conveniently presented in the statement format. The third form uses phrases that identify current issues; we feel this form is least helpful to the examinees. (Note that in each case the six topics are about equally divided between public and educational issues, which seems to us appropriate for the examinees.)

FORM 1: STATEMENTS

Read the statements below and choose one about which you have something to say. Decide in what ways you agree or disagree with that statement and write an essay in which you explain your own position on the issue. Use the underlined key words as the title for your essay.

1. Tests of basic skills should be given in the eighth grade, and students who are not minimally competent in reading, writing, and math should not be permitted to attend high school.

2. Even if it were proven that violence on TV harms children, no one has the constitutional right to tell broadcasters what they can or cannot show.

3. People who do not have children in schools should not be required to pay school taxes.

4. The best way to solve the energy crisis is simply to make prices so high that people will have to use less gasoline.

5. Many learning and discipline problems in the schools could be avoided if boys and girls were sent to separate schools after grade six.

6. It is the duty of a school to teach students how to speak and write proper English, and therefore nonstandard dialects and foreign languages should not be tolerated in the classroom.

FORM 2: QUESTIONS

Below are six questions about which there is currently a good deal of disagreement. Choose one of the questions and write an essay giving your own personal answer to it. Use the question as the title of your essay.

1. What are the essential characteristics of a good teacher?

2. What responsibility do schools have for imparting moral and ethical values to students?

3. Why is it important for a teacher to write well?

4. What is your definition of "the good life"?

5. Who do your feel is the greatest living American?

6. How are the people you know coping with inflation?

FORM 3: PHRASES

Below is a list of six controversial issues. Choose one about which you would like to write. Write an essay setting forth your personal opinions about what are the problems involved in the issue and how they should be resolved. Use the name of the issue as the title for your essay.

1. Violence in the schools.
2. Legal drinking of alcohol at age 18.
3. The energy crisis.
4. Ability grouping in schools.
5. Living together before marriage.
6. Literacy testing of high school students.

Writing additional stimulus items (topics) for assignments in any of these formats would be simple--since it can probably be safely assumed that the issues on which most of the examinees are ready to write are those being given the most attention in the news media at any particular time.

D. Physical Appearance of the Writing Examination

The test "package" for the writing examination will consist simply of a cover sheet stapled to three blank sheets of lined 8 1/2 x 11 writing paper. The cover sheet will resemble the sample on the next page. It will contain:

1. Space for whatever biographical data is desired from the examinee;
2. The instructions for writing the examination;
3. The assignment and topics;
4. A space for recording the examination code number assigned the examinee; and
5. A space for recording the score given to the examinee's essay by the raters.

Insert Sample Cover Sheet Here

E. Time to be Allowed for Writing the Examination

College Board examinations of writing ability in the early 1900's allowed students a full three hours to demonstrate their competence. More recent examinations have allowed students as little as twenty minutes to produce a sample essay. (These brief essays, though, have been supplemented by objective examinations of technical skills and knowledge). Since the essay sample on the Teacher Competency Examination will form the sole basis for judgment of an examinee's competence in writing, we would seriously question the validity of essay samples produced in so short a time period as twenty or thirty minutes.

The very brief period is especially unfair to the student who may be bright and technically competent, but not glib enough to be able to reel off his or her thoughts at high speed. We

would therefore strongly recommend that no less than forty-five minutes be provided for the writing subtest, and that, if possible, a whole hour be provided.

STATE OF FLORIDA TEACHER COMPETENCY EXAMINATION

SUBTEST OF WRITING SKILLS

Examination Code Number

NAME_____

Score

(Other information requested here as needed.)

INSTRUCTIONS. This portion of the examination gives you a chance to show how well you can write. The question below asks you to compose an essay setting forth your personal opinions or beliefs on some important issue. You should assume you are addressing your essay to an audience of educated adults. Your purpose will be to convey your position as clearly as possible to your readers.

There are, of course, no "right answers" on this examination. Your essay will be read by at least three readers and judged on its quality as a prose composition. So use your time well--plan before you begin to write, then read your essay carefully after you have finished and make any necessary corrections and revisions. The evaluation of your essay will in no way depend on whether your readers happen to agree with your opinions. (But any reader will naturally appreciate legible handwriting.)

Relax, take a deep breath, and do the best you can.

THE ASSIGNMENT. Below are six questions about which there is currently a good deal of disagreement. Choose one of the questions and write an essay giving your personal answer to it. Use the question as the title of your essay.

1. What are the essential characteristics of a good teacher?

2. What responsibility do schools have for imparting moral and ethical values to students?

3. Why is it important for a teacher to write well?

4. What is your definition of "the good life"?

5. Who do you feel is the greatest living American?

6. How are the people you know coping with inflation?

E. Comparisons Between Our Recommendations and College Board Practices

The recommendations above differ somewhat from those made, for example, by David P. Harris in an article describing the College Board's experience in trying to assess writing ability. His advice about the characteristics of test assignments is sometimes impractical or inappropriate for the purposes of the Teacher Competency Examination, as indicated in the notes below.

"1. Arrange to take several samples, rather than just one" and have them "written at different times." This procedure has been statistically demonstrated to yield the most highly reliable scores, but it would be unreasonable to ask certification candidates to appear on, say, two different weekends to write essays, particularly when many of them would be coming from out-of-state.

"2. Set writing tasks that will yield a broad range of scores." Harris' concern here is to set questions difficult enough to "encourage the very best students to perform at their full capacity." This is not a pertinent concern in the present instance, where the intention is simply to distinguish between competent and incompetent writers.

"3. Allow no alternative topics. If some students are performing different tasks from others, it is difficult to compare performances." As explained above, we have, trying to balance this consideration against the necessity of finding topics that are fair to a wide range of examinees, recommended a single, simple description of the writing task combined with an array of optional subject matters, selected with the examinees in mind.

"4. Make the writing task(s) clear and specific; provide full directions." The instructions and assignments above, we believe, satisfy these criteria.

"5. Pre-test writing-test assignments." This has been done to some extent and will be done thoroughly during the field tests of the examination.

See David P. Harris, "The Testing of Student Writing Ability," in Gary Tate, ed., Reflections on High School English, (Tulsa, OK: University of Tulsa Press, 1966), pp. 137-145.

5. ESTABLISHING A PASS/FAIL CUTOFF SCORE

A rating of "1" designates an inadequate or failing essay. A rating of "2" designates one that is minimally competent but passing. An essay that was awarded a "2" by each of the three raters for a score of "6" would, then, clearly be a passing essay. But what of an essay that two raters valued as a "2" while the third rated it as a "1"--for a score of "5"? It would be our inclination--and our recommendation--that in such a case the vote of the majority of raters be honored and that the "5" score be established as the minimal passing score.

We would further recommend, though, that in the case of a score of "4"--two raters failing the paper with a "1" and one rater passing it with a "2"--the essay should be forwarded to the referee for a fourth reading, simply to give the examinee the benefit of the doubt in the borderline case and to make the whole procedure more defensible against protests that might be registered by examinees receiving a failing grade. If the referee were to award the contested essay a rating of "2," that rating would replace one of the "1" scores and give the essay a passing grade of "5"; but if the referee gave it a "1" rating, that rating would replace the "2" rating and give the essay a clearly failing grade of "3." In effect, this procedure eliminates the possibility of an essay ending with a "barely failing" grade of "4."

6. LOCATING AND RECRUITING RATERS

A. Qualifications of Raters.

1. Technical competence. It is essential that the raters be persons who have had considerable experience in evaluating writing. It would be practically impossible to train an inexperienced group of readers up to acceptable standards of agreement within a reasonable period of time. In effect this means that the raters will be selected from among the ranks of successful high school English teachers, college composition teachers, or (possibly) professional copy editors.

2. Willingness to be trained. Persons selected as raters must be willing to be trained to follow a uniform set of procedures in rating testees' essays. It is well known that a group of equally competent and experienced readers will award vastly different valuations to a single essay if each follows his or her own personal set of criteria. In order for a group of raters to obtain the desired levels of inter-rater agreement, each rater must be willing temporarily to suppress his or her own habits and preferences and to follow a uniform set of ratings procedures.

A rater who subbornly persists in following his or her grading preferences would be a threat to the reliability of the whole ratings process and would have to be dismissed. Firing a rater would be difficult and embarrassing, so it is obviously preferable that all potential raters have explained to them before they are recruited precisely what they will be expected to do. This would allow the person who is unwilling to commit himself or herself to abandoning temporarily his or her own standards to reject the invitation to become a rater.

B. Sources of Potential Raters' Names.

It is beyond the scope of this handbook to draw up a detailed plan for locating raters. In fact any such attempt on our part would be captious, since the Department of Education has resources and procedures for locating appropriate personnel that are superior to any we can suggest. The high school English teachers who have already been involved by COTE in identifying generic competencies in English would logically be consulted both as potential raters and as nominators of other teachers who might serve as raters. School administrators and language arts supervisors, freshman composition directors in universities, and English department chairpersons in colleges and community colleges are other obvious sources of nominations. Requests for nominations of raters should stress the importance of the raters possessing similar background, and the qualifications identified above: technical competence, extensive experience, and a willingness to be trained.

C. Recruiting the Raters.

Similarly, communications to potential raters should describe the ratings process that will be engaged in and stress the fact that the success of the process depends upon the raters' willingness to commit themselves to follow a uniform set of evaluation procedures, even though the rater might dislike the procedures and find them inferior to his own preferred practices. The potential raters should be asked to reject the invitation if they feel they cannot conscientiously commit themselves to such an

agreement.

D. Selecting Referees.

The referees might be described as Master Raters. They should be the persons from within the pool of raters who have the best reputations for success as composition teachers, and who have shown and expressed the most interest in and enthusiasm for the ratings process and the whole competency program. They need not necessarily have the most years of experience. Nor should they be drawn from any particular class of raters to the exclusion of another--that is to say, high school, community college, college, and university personnel should all be represented among the referees.

7. SELECTING TRAINERS AND ADMINISTRATORS

A. Trainer Characteristics

The person who is assigned responsibility for conducting the training of the raters should have, like the raters and referees, extensive training and experience in teaching writing and evaluating essays. This background has been assumed in the writing of the Training Manual (Volume 2). Ideally, the person should also have a record of proven success in training teachers or other adults in workshop situations similar to that described in the Training Manual.

B. Administrator Characteristics

The ratings process as described in Volume 3 can be coordinated by anyone who has successful experience in supervising an operation of this order of magnitude and complexity. No special familiarity with either composition teaching or this particular kind of testing would necessarily be required. However, it may be deemed most efficient to give the same person responsibility both for the training of raters and supervision of the ratings process, since these two tasks are essentially aspects of the same operation. If the decision is to make such a unitary assignment, then it will be necessary that the administrator have the qualifications of a trainer as well as the requisite administrative expertise.

8. RELIABILITY AND RATER AGREEMENT

A. Previous Experience with Holistic Evaluation

Starch and Elliott's essay on the "Reliability of Grading High School Work in English" in a 1912 issue of School Review was the first publication on this subject. Foley's chapter on writing in the Handbook on Formative and Summative Evaluation of Student Learning (1971) summarized the contributions made to the topic since then by the CEEB (starting in 1914), Eley (1953), Huddleston (1954), Diederich (over a period of thirty years), Meckel (1963), Nyberg (1968), and Coffman and Kurfman (1968). In 1977, Cooper (in the essay already cited above) reviewed these and more recent studies.

> In 1934 a researcher demonstrated that rater reliability could be improved from a range of .30 to .75 before training to a range of .73 to .98 after training (Stalnaker, 1934)... A more recent study (Follman and Anderson, 1967) reports reliabilities for five raters ranging from .81 to .95 on five different types of holistic evaluations. Another recent study (Moslemi, 1975) reports a reliability of .95 for three raters scoring "creative" writing. In a school-district curriculum evaluation study just completed here at Buffalo, Lee Odell obtained agreements between two raters of 80%, 100%, and 100% in choosing the better essay in each of thirty pairs....
>
> As emphatically as I can, then, let me correct the record about the reliability of holistic judgments: When raters are from similar backgrounds and when they are trained with a holistic scoring guide...they can achieve...scoring reliabilities in the high eighties and

low nineties on their summed scores from multiple pieces of a student's writing. (pp. 18-19)

Cooper emphasized, however, that such high reliabilities can rarely be achieved from a rating of one paper; and Diederich (1974, cited above) gives a formula for computing how many samples of a student's writing one would have to rate in order to obtain a desired degree of reliability. As we have noted above, it is not practical to have certification candidates write on more than one occasion. And we have chosen not to recommend that they be asked to write two brief essays on the occasion of the writing subtest. There are two reasons for this. First, all authorities agree that multiple writing samples written at the same time will not demonstrate the desired variability, so there would be little gain in reliability; second, we have serious doubts about the validity of a writing sample produced in a period of twenty minutes or so.

We have, instead, striven to increase rater agreement by using a combination of three raters and a referee and by prescribing a more thorough and extensive training program--involving both detailed criteria and discussion of many sample essays--than has been used in any other program with which we are familiar.

In devising the examination specifications, the training procedures, and the ratings protocols, we have tried to apply, to the extent possible within the constraints of the given situation, the findings about causes of variation in writing performance and rating judgment that have been identified in the research literature. For further discussion of the factors related to such variations,

see Britton, Martin, and Rosen, Multiple Marking of Compositions (London: Her Majesty's Stationery Office, 1966) and McColly, "What Does Educational Research Say About the Judging of Writing Ability?" Journal of Educational Research (1970), pp. 148-56.

B. What Sort of Reliability is Appropriate?

The technical literature on reliability is voluminous and rapidly proliferating. There seem to be literally dozens of ways of computing reliability and, all too often, in the literature of essay evaluation, reliability numbers are presented without any clear specification of how they were calculated. Singleton, in an unpublished doctoral dissertation done at the University of Georgia (1976; summarized in Roberts and Rentz, cited above), compared four methods of computing reliability of scores awarded by three raters on the essay portion of the Georgia Regents' Language Skills Examination. One analysis, in which a product-moment correlation was computed between scores awarded by "expert judges" and scores awarded to the same essays during a regular rating session, yielded a correlation of .624. A second approach used Ebel's procedure for computing interclass correlation and involved analysis of variance. Reliability of average ratings was found to be .725, which Singleton interpreted to reflect "an estimate of reliability free of rater bias" and found to compare favorably with other reports of rater reliability. A third analysis involved the computation of a coefficient of concordance and yielded a reliability estimate of .821.

Singleton's fouth analysis--which resembles closely the

approach we are going to recommend in the following section as the most meaningful to the average reader--reported rating reliability in terms of "percentages of various rater-agreement patterns."

> For the 92,469 essays scored, at least two out of three raters agreed on 92.97% of the papers. Total rater agreement occurred on 34.13% of the papers. From these values, it appears that the particular procedures used in the testing program are resulting in reliable ratings. (in Roberts and Rentz, 1977., p. 30)

Before proposing a method for reporting patterns of rater agreement, we should perhaps note that many standard methods of computing rater reliability are not applicable or well-suited to the present situation for a number of reasons.

1. The writing examination is, in effect, a criterion-referenced examination, since the only basis for classification of results is whether an examinee's score is above or below the cutoff score.*

2. With only four possible ratings, there can be relatively little variability among raters (some researchers have used rating scales with as many as eight or ten gradations of quality).

---

* Hills, Gallini, and King, "Test-Retest Reliability Study of...the Statewide Assessment Test," submitted to the DOE's Bureau of Program Support Services, 1979, discusses the inappropriateness of classical reliability computations to criterion-referenced tests. A 1978 report by Brewer to the same agency reviews "Criteria-Referenced Reliability Indices" and identifies several that might be used in situations involving a single test administration.

3. The great majority of scores can be expected to be above the cutoff score (i.e., will demonstrate mastery).

4. With only a single essay sample being written, there is no subject variability, the assumption of which is basic to most reliability calculations.

In the opinion of Professor F. J. King of Florida State University, the statistical consultant to this project, these circumstances dictate that for general reporting purposes simple arithmetical computations of percentages of rater agreement are preferable.

Another sort of reliability estimate might be desired, however, for research purposes or to compare reliabilities obtained on the Florida examination with those obtained from similar projects. For these purposes, Professor King recommends the ALPHA coefficient (Cronbach's alpha). The program reference for this is David Specht, "SPSS: Statistical Package for the Social Sciences Version 6 Users Guide to Subprogram RELIABILITY and Repeated Measurements Analysis of Variance." This program is a supplement to the Statistical Package for the Social Sciences, 2nd. Edition (New York: McGraw-Hill, 1975), and is distributed through the Statistical Laboratory, Iowa State University, Ames, Iowa 50010.

C. Reporting Patterns of Rater Agreement

We propose and illustrate in this section four measures that will rather fully describe the patterns of rater agreement.

A good estimate of rater performance could be obtained from a sample of, say, ten to twenty percent of the ratings; and percentages of agreement for that number of ratings could (with only the assistance of a handheld calculator) be figured directly from the summary sheets used to record the ratings. Computing the figures for the entire population of ratings would better be done by a computer, though this would require some time-consuming preparation. The four index measures or indices are these:

1. Percentage of complete agreement among three raters. (Note that in all cases, computation is done after the referee's rating, if there is one, has replaced that of the most discrepant rater.)

2. Percentage of cases in which two out of three raters agree on a rating.

3. Average percentage of agreement between pairs of raters within a team as to passing and failing ratings. (Note that the percentage of agreement of 2 out of 3 raters as to passing or failing is by definition 100% and therefore useless as a measure of reliability.)

4. The percentage of complete agreement among raters as to passing and failing ratings.

The computation of these four measures is illustrated below using data on the ratings given to twenty essays chosen at random from among those written for our work at Florida State University. SEE TABLE 1. (These data are a sample from the ratings of three raters whose overall coefficient of reliability was .82.) A plus sign means yes, a minus, no.

TABLE 1, Sample Rater Agreement Data

Essay	Rater A	Rater B	Rater C	Index 1. Complete Agreement?	Index 2. 2 of 3 Agreeing?
1	3	3	3	+	+
2	1	1	1	+	+
3	1	2	1	-	+
4	2	2	2	+	+
5	3	4	4	-	+
6	3	2	2	-	+
7	2	3	2	-	+
8	2	1	1	-	+
9	2	2	2	+	+
10	1	1	1	+	+
11	3	4	2	-	-
12	4	2	3	-	-
13	2	3	2	-	+
14	3	3	3	+	+
15	3	3	2	-	+
16	3	2	2	-	+
17	2	1	1	-	+
18	2	2	2	+	+
19	1	1	1	+	+
20	2	1	1	-	+
				+ = 8	+ = 18

Index 1: Percent complete agreement.....40%

Index 2: Percent 2 raters agreeing.....90%

Index 3, average percentage of agreement about whether an essay should be awarded a passing or failing grade, is computed by comparing the agreements of all pairs of raters, A-B, B-C, A-C, and

dividing the summed percentages of agreement by the number of pairs. This computation for the above data is illustrated below in Table 2. A plus sign signifies agreement, a minus sign disagreement.

TABLE 2. Index 3 Agreement by Pairs about Pass/Fail

Essay	A-B	B-C	A-C
1	+	+	+
2	+	+	+
3	-	-	+
4	+	+	+
5	+	+	+
6	+	+	+
7	+	+	+
8	-	+	-
9	+	+	+
10	+	+	+
11	+	+	+
12	+	+	+
13	+	+	+
14	+	+	+
15	+	+	+
16	+	+	+
17	-	+	-
18	+	+	+
19	+	+	+
20	-	+	-
Number of Agreements	16	19	17
Percentage of Agreement	80%	95%	85%

Average percentage of agreement $\frac{80 + 95 + 85}{3} = 86.7\%$

Index 4, the percentage of complete agreement among raters as to whether a particular paper should be awarded a passing or failing rating, can be obtained by inspection of the array of ratings in Table1.

1. Only on essays 3, 8, 17, and 20 did one rater award a failing ("1") rating while another awarded a passing rating. This is sixteen cases out of twenty of complete agreement about the passing or failing status of essays, or 80%.

Of these four indices, we feel that the second and the third are the most useful for purposes of general description of patterns of rater agreement; while the third and fourth, since protests against the testing procedure will originate from failing examinees or their representatives, are perhaps the most crucial. The first index, percentage of complete agreement, is a good indicator of the success of the training, but even in the best of cases, it will be so low as to be unimpressive and subject to misunderstandings if reported publicly.

It is difficult if not impossible to predict how high each of these figures might go, or to assert how low they can fall without casting doubt on the credibility of the testing procedures. The following ranges of rater agreement on these four indices however, may serve as tentative target figures, pending actual field experience with the writing examination. (We consider these target ranges conservative, however, and would hope and expect they will be exceeded.)

TABLE 3. Target Agreement Figures

Name of Index	Target Range
1. Percentage complete agreement	30% to 40%
2. Percentage 2 of 3 agreeing	80% to 90%
3. Average percentage agreement by pairs as to passing or failing	80% to 90%
4. Percentage complete agreement as to passing or failing	70% to 80%

9. DEVELOPMENTAL USES OF DATA FROM THE FIRST ADMINISTRATION OF THE EXAMINATION

The first administration of the examination will provide materials to be used in further improving the training procedures for the raters. Specifically, it will provide sample essays written by actual certification candidates under examination conditions and rated by raters trained according to the specifications in these volumes. A selection of these essays, chosen to represent the range and variety of ratings and rating problems, should replace the sample essays now contained in the Training Manual (which were written by upperclassmen at a single university and rated by raters who had undergone a similar but less intensive training program). Such a replacement should make the training task resemble even more closely the actual ratings task--a minor change, admittedly, but one which might contribute at least a bit to the improvement of rater reliability.

In a more general way, all the experience gained in the course of field testing and actually administering this examination

should be utilized to improve the effectiveness and efficiency of the examination.

10. POSSIBLE RESEARCH USES OF THE EXAMINATION ESSAYS AND RATINGS

Each administration of the writing examination will yield a rich corpus of data that should be open to researchers interested in investigating such problems as the following:

A. What are the relationships between examinee characteristics (e.g., sex, institution awarding degree, ethnicity, major field of study, etc.) and scores on the writing examination?

B. What are the correlations (if any) between characteristics of essays (e.g., length, rhetorical mode, vocabulary, "syntactic maturity," etc.) and

1. ratings awarded, or
2. rater agreement as to ratings?

C. What is the biographical "profile" of examinees receiving failing grades? outstanding grades?

D. What relationships (if any) exist between the topic an examinee chooses to write upon and the score he or she receives?

E. What can analysis of essays written for the examination reveal about common examinee weaknesses in writing and test-taking ability that should be addressed by college programs?

F. Are teachers who make extremely high scores on this examination more successful in their first year of teaching than those who make barely passing scores? In what ways and why?

The list of possible important topics could be extended indefinitely. The point to be made is simply that this examination will not only serve to screen teachers, it will also produce a great mass of data which may be used both to further our understanding of some of the issues involved in such an examination and to provide information that may be used to improve the teacher preparation programs in our colleges.

11. SUMMARY OF MAJOR RECOMMENDATIONS

A. That the form of holistic evaluation to be used is "general impression marking";

B. That the assignment format consist of a single set of instructions followed by at least six optional topics.

C. That the examinees be given a minimum of forty-five minutes to write the sample essay, with a full hour being provided if possible.

D. That the training process include both criteria and extensive reading of graded sample essays.

E. That teams of three raters be used to score essays, with a fourth reader or referee to be used to reconcile discrepant scores.

F. That raters, referees, and trainers all have backgrounds as writing teachers in high schools, colleges, or universities;

G. That the cutoff score be "5" on a scale running from "3" to "12";

H. That reporting of reliability be done in terms of various indices of rater agreement.